Published online 11 June 2012 *Nucleic Acids Research, 2012, Vol. 40, Web Server issue* ***W173–W179***
doi:10.1093/nar/gks564

DBD2BS: connecting a DNA-binding protein with its binding sites

Ting-Ying Chien[1], Chih-Kang Lin[2], Chih-Wei Lin[1], Yi-Zhong Weng[1], Chien-Yu Chen[2,3,*] and Darby Tien-Hao Chang[4,*]

[1]Department of Computer Science and Information Engineering, [2]Center for Systems Biology, [3]Department of Bio-Industrial Mechatronics Engineering, National Taiwan University, Taipei 106, Taiwan, and [4]Department of Electrical Engineering, National Cheng Kung University, Tainan 701, Taiwan

Received March 4, 2012; Revised May 7, 2012; Accepted May 19, 2012

ABSTRACT

By binding to short and highly conserved DNA sequences in genomes, DNA-binding proteins initiate, enhance or repress biological processes. Accurately identifying such binding sites, often represented by position weight matrices (PWMs), is an important step in understanding the control mechanisms of cells. When given coordinates of a DNA-binding domain (DBD) bound with DNA, a potential function can be used to estimate the change of binding affinity after base substitutions, where the changes can be summarized as a PWM. This technique provides an effective alternative when the chromatin immunoprecipitation data are unavailable for PWM inference. To facilitate the procedure of predicting PWMs based on protein–DNA complexes or even structures of the unbound state, the web server, DBD2BS, is presented in this study. The DBD2BS uses an atom-level knowledge-based potential function to predict PWMs characterizing the sequences to which the query DBD structure can bind. For unbound queries, a list of 1066 DBD–DNA complexes (including 1813 protein chains) is compiled for use as templates for synthesizing bound structures. The DBD2BS provides users with an easy-to-use interface for visualizing the PWMs predicted based on different templates and the spatial relationships of the query protein, the DBDs and the DNAs. The DBD2BS is the first attempt to predict PWMs of DBDs from unbound structures rather than from bound ones. This approach increases the number of existing protein structures that can be exploited when analyzing protein–DNA interactions. In a recent study, the authors showed that the kernel adopted by the DBD2BS can generate PWMs consistent with those obtained from the experimental data. The use of DBD2BS to predict PWMs can be incorporated with sequence-based methods to discover binding sites in genome-wide studies. Available at: http://dbd2bs.csie.ntu.edu.tw/, http://dbd2bs.csbb.ntu.edu.tw/, and http://dbd2bs.ee.ncku.edu.tw.

INTRODUCTION

The DNA-binding proteins use DNA-binding domains (DBDs) to recognize specific nucleotide sequences in genomes. Such short and highly conserved DNA sequences are usually summarized by a probabilistic model called a position weight matrix (PWM) (1–3). Accurately constructing PWMs for DBDs is an important step in the understanding of many biological processes. Recently, successful incorporation of inchromatin immunoprecipitation (ChIP) with tiling arrays or next-generation sequencing enabled discovery of the binding locations of DNA-binding proteins in genomes after invoking motif finding tools (4–8). However, a limitation of ChIP experiments is their inability to distinguish direct binding from indirect interactions. Therefore, an auxiliary method of analyzing physical protein–DNA interactions at atomic level is urgently needed.

Many studies show that PWMs can be inferred from DBD–DNA complex structures with a satisfactory accuracy (9–11). In this task, potential functions play a key role in identifying nucleotides with high specificities. Existing potential functions for protein–DNA interactions are roughly categorized as physics-based (12,13) and knowledge-based (11,14,15). Physics-based potential functions use empirical components derived from physics, including electrostatics, hydrogen-bond and van der Waals force. Knowledge-based potential functions, on the other hand, adopt statistical components, such as the

*To whom correspondence should be addressed. Tel: +886 2 33665334; Fax: +886 2 23627620; Email: cychen@mars.csie.ntu.edu.tw
Correspondence may also be addressed to Darby Tien-Hao Chang. Email: darby@ee.ncku.edu.tw

number of contacts and the distance distribution between the contacts, derived from known protein–DNA structures. For PWM inference, knowledge-based potentials that consider all atom types show prediction accuracy comparable with that of physics-based potentials but at much lower computation cost (11).

Our recent study demonstrated that the same idea can be extended to unbound DBDs (DBD structures of the unbound state) (16). The 26 December 2011 release of Protein Data Bank (PDB) (17) contains the structures of 1373 DNA-binding proteins, where <500 proteins have protein–DNA co-crystallized structures. This reveals an immediate need to develop PWM predictions for unbound structures. Gao and Skolnick recently proposed an efficient way of using structure alignment with a template to generate protein–DNA complexes from structures of unbound proteins (18). In this study, the template is a protein–DNA complex with DBD folds similar to those of the unbound structure. The complexes generated by structure alignment are shown reliable in predicting PWMs consistent with the experimental data (16), providing a more efficient alternative to docking (19,20) and a more accurate alternative to homology modeling (10).

By providing an automatic and integrated platform for these procedures, this work helps researchers analyze protein–DNA interactions. A list of 1066 DBD–DNA complexes (including 1813 protein chains) is compiled for use as the template database. For a given DBD–DNA complex, DBD2BS uses an atom-level knowledge-based potential function to infer PWMs. For protein structures without existing co-crystallized complexes, DBD2BS conducts structure alignment to synthesize the bound state of the query structure and then performs PWM prediction based on the synthetic DBD–DNA complexes. The DBD2BS is the first attempt to predict PWMs of DBDs from unbound structures rather than from bound ones. Using unbound structures increases the number of existing protein structures that can be exploited for analyzing protein–DNA interactions. The DBD2BS also provides users with an easy-to-use interface for visualizing the PWMs predicted based on different templates and the spatial relationships of the query protein, the DBDs and the DNAs.

MATERIALS AND METHODS

Figure 1 shows the workflow of DBD2BS, where the query could be a DBD structure or a DBD–DNA complex structure. If the query is a DBD structure (Figure 1a), DBD2BS searches the template database and generates appropriate DBD–DNA complexes for PWM prediction. This section first describes the construction of the template database. Next, the structure alignment performed to generate appropriate complexes for the query is described. The final section describes the use of the all-atom knowledge-based potential function for predicting PWM.

Constructing templates

The DBD2BS templates are built based on the protein–DNA complex structures collected from PDB. A complex is selected as a template if (i) it is an X-ray structure with resolution better than 3.0 Å, (ii) it contains exactly one double-strand DNA, (iii) the DNA molecule has six or more paired bases and has <30% non-paired bases,

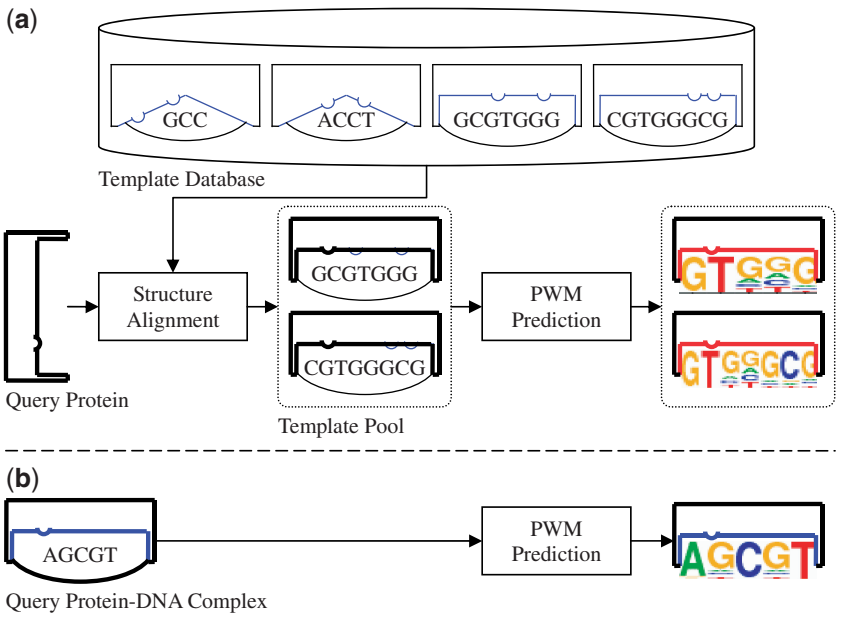

Figure 1. Prediction workflow incorporated in DBD2BS. (**a**) If the query is a DBD structure, DBD2BS searches the template database and uses structure alignment to generate appropriate DBD–DNA complex structures (the template pool) for PWM prediction. The DBD2BS reports both the potential DBDs on the query protein and the corresponding PWMs. The red binding interfaces indicate the potential DBDs. (**b**) If the query is a DBD–DNA complex, no extra action is required before PWM prediction. The DBD2BS reports the predicted PWMs, where the blue binding interface indicates that the DBD is known on the query rather than predicted.

(iv) at least one protein chain has five or more contact residues (residues within 4.5 Å to the DNA molecule) and (v) at least one protein chain has ≥40 residues. Based on the PDB release of 26 December 2011, a template collection of 1066 protein–DNA complexes containing 1813 protein chains was constructed.

Finding appropriate DBD templates

To improve efficiency of the service, a two-stage strategy of structure alignment was used to find appropriate template complexes. The collected protein chains were first clustered by the HomoClust (21) hierarchical clustering algorithm based on pair-wise structure similarities (denoted TM-scores) reported by TM-align (22). In each cluster, the protein chain with the largest average similarity to the other protein chains is assigned as the representative of the cluster. When a query protein is given, it is compared with all cluster representatives ('Structure Alignment' in Figure 1). If the TM-score is >0.5 between the query structure and the cluster representative, all the other protein chains belonging to that cluster are aligned with the query structure. Finally, only the templates with an alignment score to the query structure >0.5 are collected in the template pool. If the template pool contains less than three templates, DBD2BS finds the templates with the next highest alignment scores so that the pool contains at least three templates. These templates are then listed in the intermediate page for selection (see WEB INTERFACE section).

Figure 1 shows that the query protein is superimposed onto the template structure, where the superimposed complexes are stored in the 'Template Pool'. The superimposition is performed by applying the rotation matrix reported by TM-align. After removing the original protein chains in the template and appending the transformed coordinates of the query structure into the template structure, the DBD2BS generates a superimposed complex for PWM prediction. For visualization on the results page (see WEB INTERFACE section), DBD2BS also prepares a combined superimposed complex that keeps the ATOM coordinates of the template structure in MODEL 1 as well as the transformed ATOM coordinates of the query structure in MODEL 2.

Potential function for PWM prediction

To estimate the binding free energy of a DBD–DNA complex, the DBD2BS implements a variation of the potential function FIRE described by (11). FIRE is a succinct knowledge-based potential function that considers interactions among all atom types. A pair of atom types (one from proteins and one from DNA) is scored according to the observed frequency relative to the expected frequency estimated based on a collection of protein–DNA complexes. Accordingly, the frequency of all the combinations of atom types with respect to different scales of distances in the collection of protein–DNA complexes is recorded. After the potential between atom types a and b falling within a particular distance r, $u^{FIRE}(a,b,r)$, is scored the binding free energy of a protein–DNA complex can be statistically estimated as follows:

$$\Delta G = \sum_{a,b} u^{FIRE}(a,b,r). \quad (1)$$

Assuming that the influences of different positions on binding free energy are independent, ΔG can be represented as follows:

$$\Delta G = \sum_{j} \Delta G_{\alpha}^{j}, \quad (2)$$

where ΔG_{α}^{j} is the binding free energy contributed by a base α (A, T, C or G) at position j. Based on equation (2), the DBD2BS replaces each base in the protein-DNA complex by three alternative nucleotides to calculate the relative propensities of the four types of bases. For a more detailed description of the PWM calculation, readers can refer to (16).

WEB INTERFACE

The input page of DBD2BS includes two submission forms corresponding to the two flows in Figure 1. The first prediction mode includes an extra step for generating the DBD–DNA complex structures by performing structure alignment. In each mode, users can specify the query structure in three different ways: (i) provide a PDB ID (the protein chain ID is additionally required for the mode 'Query with a protein structure'); (ii) specify the atomic coordinates of the query structure in PDB format in the text field or (iii) upload a structure file in PDB format.

The first form 'Query with a protein structure' allows only one monomer per run. If the query contains more than one protein chain, only the first protein chain is used. After pressing the submit button, the user is directed to the template selection page, where candidate templates are sorted by the TM-scores between the template and the query structure. This page includes the cluster information, template IDs (PDB IDs), structure alignment scores, sequence alignment scores (e-values between the template sequence and query sequence by performing BLAST with default parameters), template proteins (recognized by UniProt (23) entry name) and template descriptions. Although templates with higher TM-scores are generally preferred, those with intermediate TM-scores but significant e-values can also be considered. Users are warned if the protein chain in the template is a long α-helix. The superimposed complexes may be unreliable in such cases. To prevent that the query protein in the superimposed complex has serious collisions against the DNA structure, synthesized complexes containing more than five conflicting residues are excluded by DBD2BS automatically, where a conflicting residue is defined as a residue with at least one heavy atom within 1.5 Å to any heavy atom of the DNA. Users can decide whether the predictions are based on a set of similar templates or on a more diverse set. In each run, users must select one to four templates to make PWM prediction. Users can return to the page at any time to select other templates.

The left side of the result page (Figure 2a) lists the templates selected in the previous page. The predicted PWMs are shown in the sequence logo form. Clicking the '3D' buttons of sequence logos loads the corresponding templates into the Jmol (available at http://www.jmol.org/) panel on the right side of the result page. In this panel, DBDs are displayed as 'sticks'. Users can click the 'Both' radio button to see how the query protein and the template are superimposed. The DNA base pairs are colored according to their conservation level. The conservation score is derived by calculating the position entropy in the predicted PWM. The 5′ end of the sequence logo (the position '1' in the sequence logo) in the Jmol panel is highlighted by showing the corresponding base in green so that users can quickly link the sequence logo with the DNA in the Jmol panel.

Two advanced functions are provided by DBD2BS to help users determine the reliability of the predicted PWM might be. First, atom collisions, the red sticks in Figure 2a, are highlighted by clicking the 'On' radio button above the 3D viewer. Users should be aware of potential false predictions on the base pairs close to any atom collisions. Second, for any PWM of interest, users can click the 'CMP' button of the sequence logo to see whether the selected PWM (or some of the positions) is supported by the predicted PWMs from other templates. Figure 2b

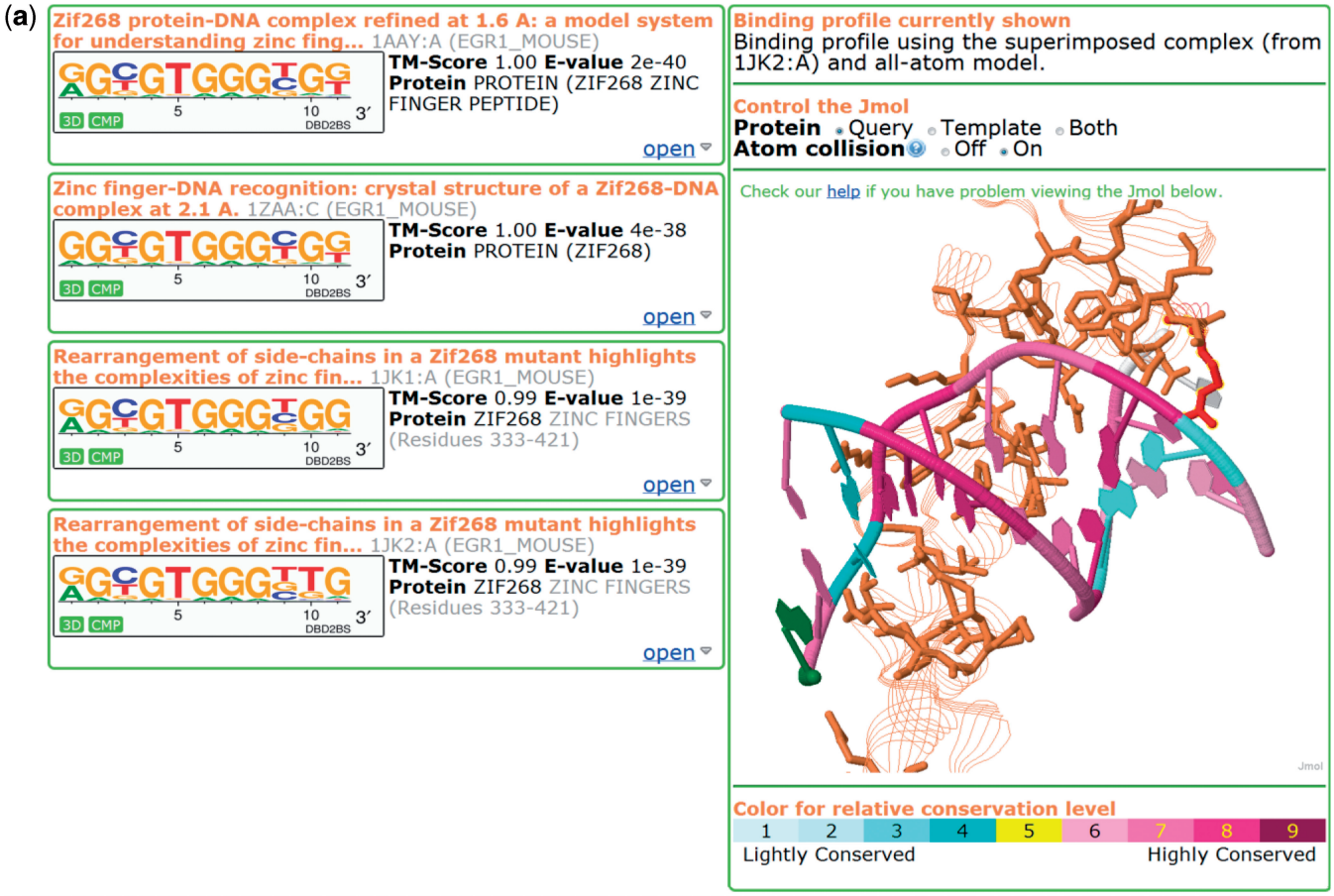

Figure 2. Screenshots of DBD2BS results. (**a**) Result page and (**b**) comparison page.

shows that the sequence logo of the selected template (3N4M:A) is highlighted on the comparison page as the reference PWM. The reference complex is first aligned to each of the other complexes by superimposing the query protein inside them. After superimposition, the DNA structures from two synthetic complexes are structurally aligned by invoking dynamic programming. Base pairs from different complexes are aligned if their distance is within 2 Å. This may result in discontinuous alignment of the sequence logos. Figure 2b shows that the unaligned positions are trimmed to produce new sequence logos. Comparing the sequence logos from different templates shows which positions have higher confidence when consistent predictions are observed.

On the comparison page, the DBD2BS also provides quantitative scores between each pair of aligned sequence logos. The similarity score with respect to a particular position from the two aligned PWMs, p and q, is defined as follows:

$$sim(p_j,q_j) = 1 - \frac{1}{\sqrt{2}}\sqrt{\sum_{\alpha\in\{A,T,C,G\}}[p_j(\alpha) - q_j(\alpha)]^2}, \quad (3)$$

where $p_j(\alpha)$ and $q_j(\alpha)$ are the frequencies of the base type α at the position j in PWMs p and q, respectively. With sim(p_j, q_j), the similarity score between p and q is defined as

$$sim(p,q) = \frac{1}{w}\sum_{j=1}^{w}\left[sim(p_j,q_j)\right], \quad (4)$$

where w is the alignment length of the two PWMs. Furthermore, the complete-similarity developed in the study of Tanaka *et al.* (24) can be calculated by:

$$sim_c(p,q) = \frac{1}{w}\sum_{j=1}^{w}\left[sim(p_j,q_j) - m_j\right], \quad (5)$$

where m_j is the median score of sim(p_j,r_j) with $r_j(\alpha)$ the frequency of the base type α at the position j in a randomly generated PWM r. When the DBD2BS uses the complete-similarity score to measure the similarity between two predicted PWMs, 10 000 randomly generated PWMs were produced to get the median score m_j.

Clicking the 'open' button in the main result page (Figure 2a) of each sequence logo reveals additional details of the selected template and its prediction results. Five files are prepared for download in this panel, including (i) the superimposed complex structure containing the query structure, (ii) the native complex structure of the template, (iii) alignment, (iv) contact residues and (v) PWM.

The second form 'Query with a protein–DNA complex' in the input page accepts a protein–DNA complex as the query, which enables users to generate protein–DNA complexes using other techniques such as protein–DNA docking (20) or homology modeling (25). This mode skips the template selection step and directly guides users to the result page. The given complex is regarded as the only one entry in the template list.

EVALUATION AND CASE STUDIES

The method adopted by DBD2BS has been demonstrated in (16) that it is able to provide reliable predictions for native protein–DNA complexes. However, (16) also showed that performance may degrade in unbound queries if conformational changes occur upon DNA binding. Therefore, users are encouraged to compare predictions using different templates for figuring out the positions in the predicted PWMs with high confidence. Furthermore, if the users have an alternative way that can generate a near-native complex, they can choose the second form of DBD2BS to upload their own synthetic complexes. In the subsequent sections, we use two cases to show the scenarios of (i) using an unbound query with the comparison facility of DBD2BS and (ii) using a bound query to simulate a near-native complex synthesized outside DBD2BS.

Querying DBD2BS with an unbound structure

This case uses a DNA-binding protein, catabolite gene activator in *Escherichia coli* (UniProt entry name CRP_ECOLI), as an example to illustrate how DBD2BS can facilitate the study of protein–DNA interactions and how the utilities embedded in DBD2BS can help the users to elucidate DBD2BS's output. This query protein was selected from the seven testing proteins used in (16) by the following procedures. We queried the seven protein names in the AH-DB (26), a database of protein structure pairs before and after binding, and the query structure (1G6N) turned out to be the unbound structure with the largest root-mean-square deviation (RMSD) of 4.54 to the bound structure upon DNA binding. Namely, this query protein undergoes a large conformational change from its unbound state to the bound one.

We used the 'Query with a protein structure' form to analyze the query (PDB ID: 1G6N). The four default templates suggested by DBD2BS, with the highest TM-scores, were selected for PWM prediction. The predicted PWMs are shown in Figure 3a, compared with the annotated PWM (Figure 3b) collected from (9). At the first glance, the four predicted PWMs are quite different. According to the reported TM-scores and e-values, users might infer that the last two predictions are relatively unreliable. This argument could be further confirmed by clicking the '3D' button of each PWM to load the combined superimposed complex in the 3D view. The loaded complex contains the protein–DNA template and the superimposed query to help users observe the query–DNA interactions and the conformational changes between the unbound state (the query) and the bound state (the protein in the template). In this example, although the query was superimposed well on all templates, we found that the query in the first two templates have more amino acid–base interactions than the last two templates. This implies that the first two predicted PWMs are more reliable.

To further compare the first two predicted PWMs with each other, DBD2BS provides a comparison facility to help human eyes. Figure 2b shows the comparison results by clicking the 'CMP' button of the first

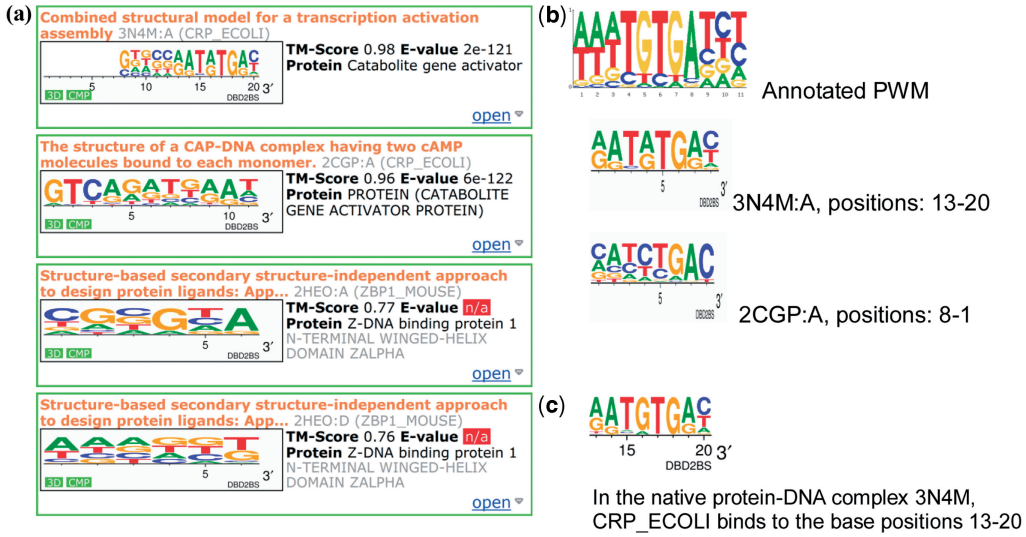

Figure 3. Case study using the catabolite gene activator of *Escherichia coli*.

template. In this figure, we found that the first two predicted PWMs had a highly consistent sub-region after reversely complementing the second one. The trimmed PWMs of the two predicted PWMs and the corresponding annotated PWM are also listed in Figure 3b. Within this eight-base region, both PWMs have at least five positions of which the most favorable base is consistent with the annotated PWM. More importantly, all the five positions were consistent in the two predicted PWMs. The last position is the only incorrect position (the most favorable base is inconsistent with the annotated PWM) among the six positions agreed by the two predicted PWMs. Finally, we moved back to the 3D view to examine the spatial characteristics of the incorrect position. It was observed that collisions happened in both superimposed structures near the incorrect position. Through this example, we summarize several guidelines of elucidating the DBD2BS's output using unbound queries: (i) use templates with higher TM-scores, which means less conformational changes; (ii) positions agreed by multiple PWMs are more reliable; (iii) boundary positions are less reliable because of fewer protein–DNA contacts and (iv) positions near collisions, which indicate large conformational change, are less reliable.

Querying DBD2BS with a protein–DNA complex

The efficiency of structure alignment comes from that it regards proteins as rigid bodies. In the previous case, this strategy leads to some coordinate differences of the synthetic complexes to the native ones. If the users have an alternative way that can generate a near-native complex, they can choose the second form of DBD2BS to upload their own synthetic complexes. In this study, we used a native complex of CRP_ECOLI (PDB ID: 3N4M) to simulate such a situation. The result is shown in Figure 3c. Among the eight bases of the positions 13–20, the predicted PWM based on the near-native complex achieves seven correct bases. This is better than the conclusion of analyzing multiple predicted PWMs in the previous case. This echoes with (16) about the importance of conformational changes in PWM prediction and encourages users to produce synthetic protein–DNA complexes through other time-consuming offline methods such as molecular docking.

CONCLUSION AND FUTURE PERSPECTIVES

Accurate construction of binding profiles for DNA-binding proteins is an important step for studying protein–DNA interactions. This article proposes the web server DBD2BS for predicting PWMs for protein–DNA complexes and unbound protein structures. We demonstrated how an unbound protein structure can use existing native protein–DNA complexes in PDB to predict its own binding sites by using DBD2BS. The PWMs predicted by DBD2BS can be incorporated with other sequence-based methods to discover more binding sites in the near future.

ACKNOWLEDGEMENTS

The authors would like to thank National Science Council of Taiwan and the Center for Systems Biology, National Taiwan University for the financial support.

FUNDING

National Science Council of Taiwan [NSC 100-2627-B-002-002] and the Center for Systems Biology, National Taiwan University. Funding for open access charge: National Science Council of Taiwan.

Conflict of interest statement. None declared.

REFERENCES

1. Bulyk,M.L. (2003) Computational prediction of transcription-factor binding site locations. *Genome Biol.*, **5**, 201.
2. Stormo,G.D. (2000) DNA binding sites: representation and discovery. *Bioinformatics*, **16**, 16–23.
3. Siggia,E.D. (2005) Computational methods for transcriptional regulation. *Curr. Opin. Genet. Dev.*, **15**, 214–221.
4. Sandelin,A. and Wasserman,W.W. (2004) Constrained binding site diversity within families of transcription factors enhances pattern discovery bioinformatics. *J. Mol. Biol.*, **338**, 207–215.
5. Xing,E.P. and Karp,R.M. (2004) MotifPrototyper: a Bayesian profile model for motif families. *Proc. Natl Acad. Sci. USA.*, **101**, 10523–10528.
6. Mahony,S., Golden,A., Smith,T.J. and Benos,P.V. (2005) Improved detection of DNA motifs using a self-organized clustering of familial binding profiles. *Bioinformatics*, **21(Suppl 1)**, i283–i291.
7. Mahony,S., Hendrix,D., Golden,A., Smith,T.J. and Rokhsar,D.S. (2005) Transcription factor binding site identification using the self-organizing map. *Bioinformatics*, **21**, 1807–1814.
8. Macisaac,K.D., Gordon,D.B., Nekludova,L., Odom,D.T., Schreiber,J., Gifford,D.K., Young,R.A. and Fraenkel,E. (2006) A hypothesis-based approach for identifying the binding specificity of regulatory proteins from chromatin immunoprecipitation data. *Bioinformatics*, **22**, 423–429.
9. Morozov,A.V., Havranek,J.J., Baker,D. and Siggia,E.D. (2005) Protein–DNA binding specificity predictions with structural models. *Nucleic Acids Res.*, **33**, 5781–5798.
10. Morozov,A.V. and Siggia,E.D. (2007) Connecting protein structure with predictions of regulatory sites. *Proc. Natl Acad. Sci. USA*, **104**, 7068–7073.
11. Xu,B., Yang,Y., Liang,H. and Zhou,Y. (2009) An all-atom knowledge-based energy function for protein-DNA threading, docking decoy discrimination, and prediction of transcription-factor binding profiles. *Proteins*, **76**, 718–730.
12. Donald,J.E., Chen,W.W. and Shakhnovich,E.I. (2007) Energetics of protein–DNA interactions. *Nucleic Acids Res.*, **35**, 1039–1047.
13. Endres,R.G., Schulthess,T.C. and Wingreen,N.S. (2004) Toward an atomistic model for predicting transcription-factor binding sites. *Proteins*, **57**, 262–268.
14. Liu,Z., Mao,F., Guo,J.T., Yan,B., Wang,P., Qu,Y. and Xu,Y. (2005) Quantitative evaluation of protein–DNA interactions using an optimized knowledge-based potential. *Nucleic Acids Res.*, **33**, 546–558.
15. Zhang,C., Liu,S., Zhu,Q. and Zhou,Y. (2005) A knowledge-based energy function for protein–ligand, protein–protein, and protein–DNA complexes. *J. Med. Chem.*, **48**, 2325–2335.
16. Chen,C.Y., Chien,T.Y., Lin,C.K., Lin,C.W., Weng,Y.Z. and Chang,D.T. (2012) Predicting target DNA sequences of DNA-binding proteins based on unbound structures. *PloS one*, **7**, e30446.
17. Kirchmair,J., Markt,P., Distinto,S., Schuster,D., Spitzer,G.M., Liedl,K.R., Langer,T. and Wolber,G. (2008) The Protein Data Bank (PDB), its related services and software tools as key components for in silico guided drug discovery. *J. Med. Chem.*, **51**, 7021–7040.
18. Gao,M. and Skolnick,J. (2008) DBD-Hunter: a knowledge-based method for the prediction of DNA–protein interactions. *Nucleic Acids Res.*, **36**, 3978–3992.
19. van Dijk,M., van Dijk,A.D., Hsu,V., Boelens,R. and Bonvin,A.M. (2006) Information-driven protein–DNA docking using HADDOCK: it is a matter of flexibility. *Nucleic Acids Res.*, **34**, 3317–3325.
20. Liu,Z., Guo,J.T., Li,T. and Xu,Y. (2008) Structure-based prediction of transcription factor binding sites using a protein–DNA docking approach. *Proteins*, **72**, 1114–1124.
21. Chen,C.-Y., Chung,W.-C. and Su,C.-T. (2006) Exploiting homogeneity in protein sequence clusters for construction of protein family hierarchies. *Pattern Recogn.*, **39**, 2356–2369.
22. Zhang,Y. and Skolnick,J. (2005) TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.*, **33**, 2302–2309.
23. Kirchmair,J., Markt,P., Distinto,S., Schuster,D., Spitzer,G.M., Liedl,K.R., Langer,T. and Wolber,G. (2010) The Universal Protein Resource (UniProt) in 2010. *Nucleic Acids Res.*, **38**, D142–D148.
24. Tanaka,E., Bailey,T., Grant,C.E., Noble,W.S. and Keich,U. (2011) Improved similarity scores for comparing motifs. *Bioinformatics*, **27**, 1603–1609.
25. Contreras-Moreira,B., Branger,P.A. and Collado-Vides,J. (2007) TFmodeller: comparative modelling of protein–DNA complexes. *Bioinformatics*, **23**, 1694–1696.
26. Chang,D.T., Yao,T.J., Fan,C.Y., Chiang,C.Y. and Bai,Y.H. (2012) AH-DB: collecting protein structure pairs before and after binding. *Nucleic Acids Res.*, **40**, D472–D478.